TIME-DOMAIN ASTRONOMY

Lectures 8: Non Parametric Analysis

Stefano Covino
INAF / Brera Astronomical Observatory

Phase Dispersion Minimization

- It is easy to realize that a possible procedure to choose between different periods for a time series could be based on computing which is the one producing the least observational scatter about the mean (derived) light curve.

- Let's assume to have a discrete set of N observations represented by an observation time t and a magnitude x.

- And be σ^2 the sample variance of x:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N - 1},$$

Phase Dispersion Minimization

- Suppose we have chosen M distinct subsamples, having variances s_j^2 (j=1,M) and containing n_j data points. The overall variance for all the samples is then given by:

$$s^2 = \frac{\sum (n_j - 1) s_j^2}{\sum n_j - M}$$

- In order to minimize the variance of the data with respect to the mean light curve be Π a trial period and compute the phases:

$$\phi_i = t_i / \Pi - \mathrm{int}(t_i / \Pi)$$

Phase Dispersion Minimization

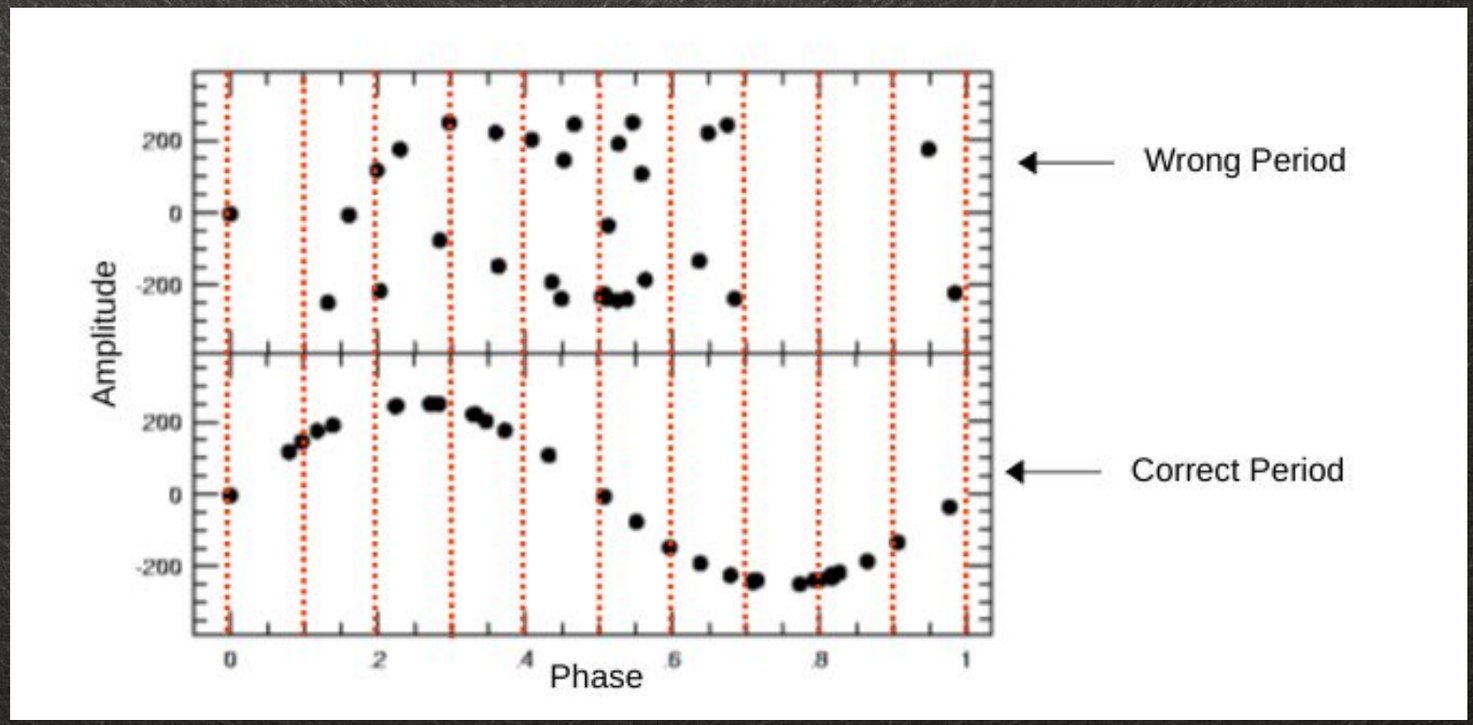
- Now let's pick M samples from x using the criterion that all the members of sample j have similar φ_j.
- The samples may be chosen in any way that satisfies the criterion. All points need not be picked, or, alternatively, a point can belong to many samples.
- The variance of these samples gives a measure of the scatter around the mean light curve defined by the means of the x_j in each sample, considered as a function of φ.

Phase Dispersion Minimization

- Finally, let's define the statistics:

$$\Theta^2 = s^2 / \sigma^2$$

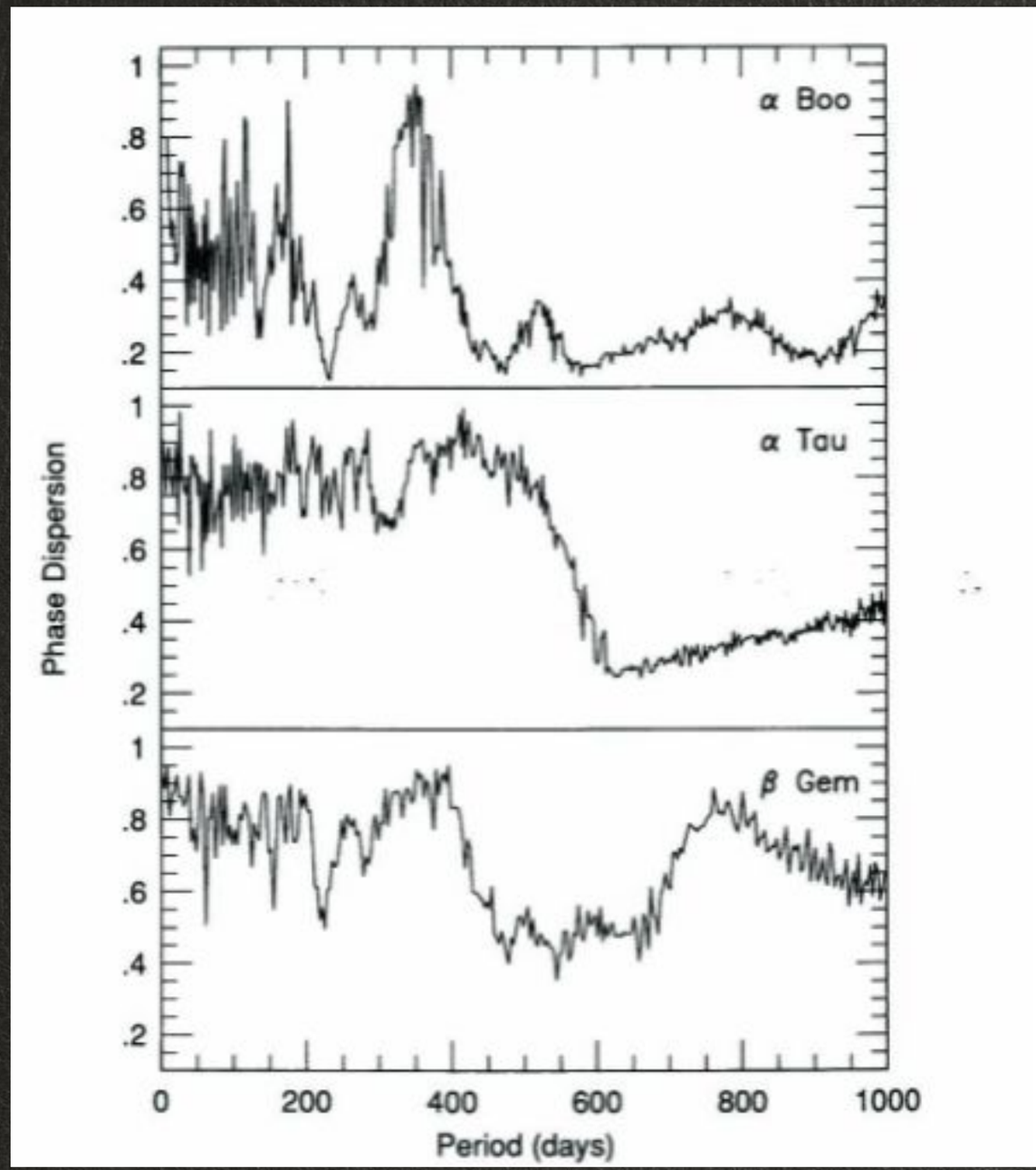
- If Π is not the true period, then $s^2 \simeq \sigma^2$, while if Π is the true period Θ will be in a local minimum compared to other tested periods.



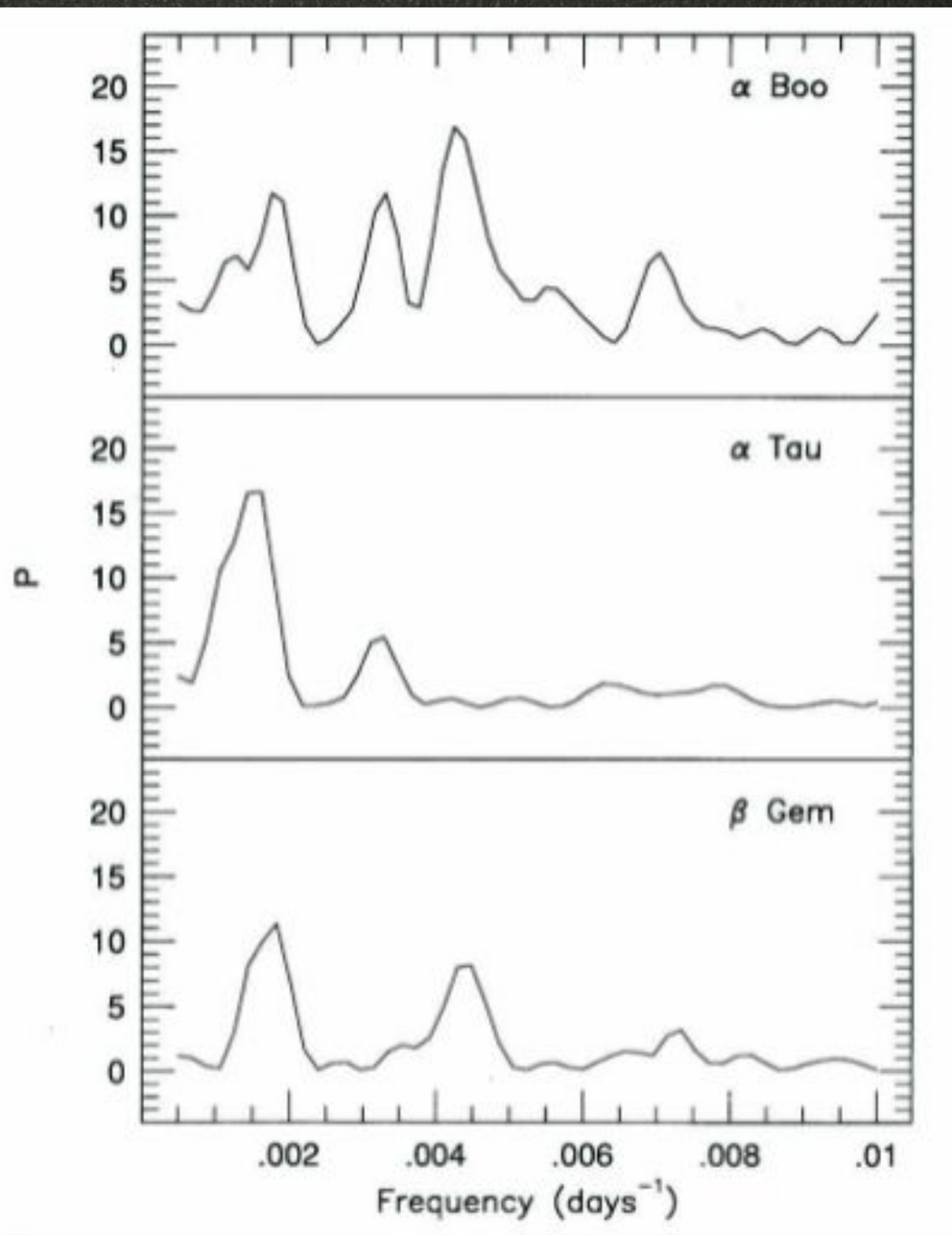
Phase Dispersion Minimization

- Since this technique seeks to minimize the dispersion of the data at constant phase, it is referred as "phase dispersion minimization" (PDM).
- If we assume that our data are normally distributed it can be shown that Θ follow a F distribution with $\sum n_j - M$ and N-1 degrees of freedom (since indeed $\Theta = s^2/\sigma^2$).

PDM vs Lomb-Scargle



PDM



Lomb-Scargle

String Length Method

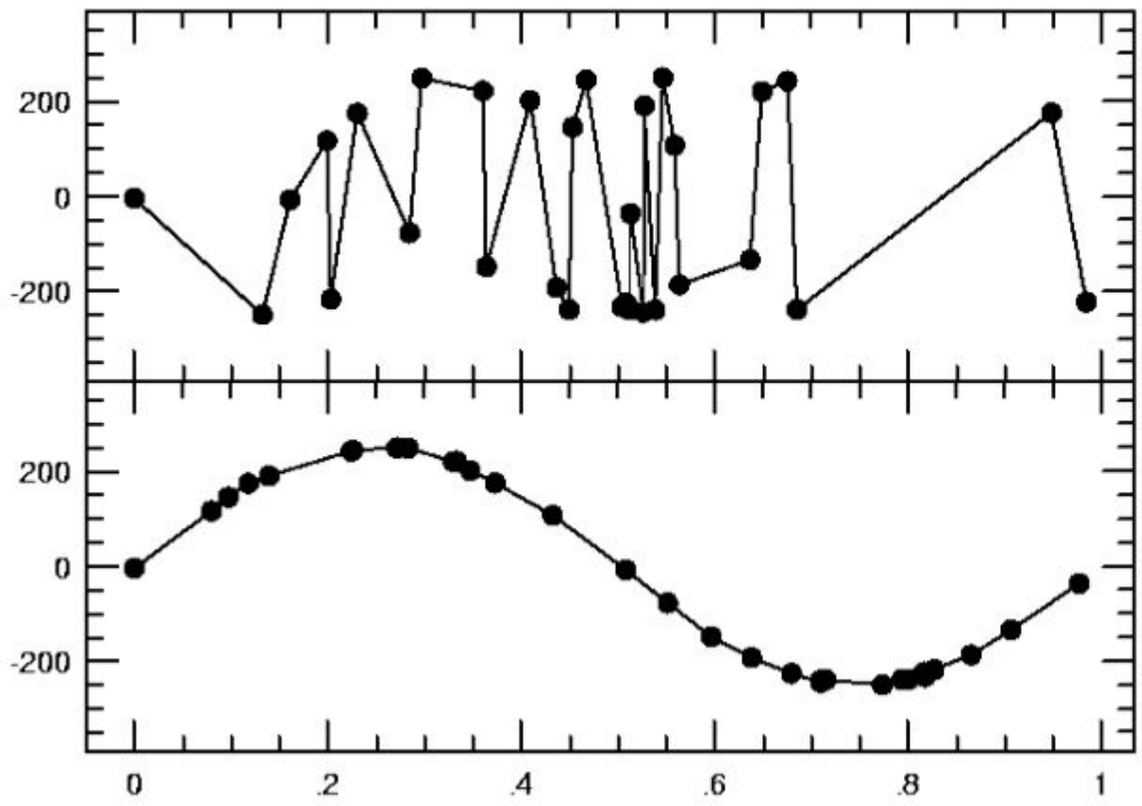
- In this method the quantity to be minimized is the sum of the lengths of line segments joining successive points (m_i, φ_i) in a phase diagram.
- The period is chosen to minimize the following quantity, with n the number of observations:

$$\sum_{i=1}^{n-1} [(m_i - m_{i-1})^2 + (\phi_i - \phi_{i-1})^2]^{1/2} + [(m_1 - m_n)^2 + (\phi_1 - \phi_n + 1)^2]^{1/2}$$

- Given that m_i and φ_i have different units one has to properly scale m_i.

String Length Method

- In its essence the algorithm is straightforward:



Wrong period longer string length

A rich zoo...

- People developed several methods to cope with the characteristics of light curves in astronomy (irregular sampling, etc.).
- The most widely used are the Lomb-Scargle (LS) periodogram, epoch folding, analysis of variance (AoV), string length (SL) methods, and the discrete or slotted autocorrelation.
- For the LS and AoV periodograms statistical confidence measures have been developed to assess periodicity detection besides estimating the period.

Lomb-Scargle

- As we know already, the LS periodogram is an extension of the conventional periodogram for unevenly sampled time series.
  - A sample power spectrum is obtained by fitting a trigonometric model in a least squares sense over the available randomly sampled data points.
  - The maximum of the LS power spectrum corresponds to the angular frequency whose model best fits the time series.

Epoch Folding

- In epoch folding a trial period, P, is used to obtain a phase diagram of the light curve.
  - The trial period is found by an ad-hoc method, or simply corresponds to a sweep among a range of values.

- This transformation is equivalent to dividing the light curve in segments of length P and then plotting the segments one on top of another, hence folding the light curve.
  - If the true period is used to fold the light curve, the periodic shape will be clearly seen in the phase diagram.
  - If a wrong period is used instead, the phase diagram won't show a clear structure and it will look like noise.

Analysis of Variance and String Length

- In AoV the folded light curve is binned and the ratio of the within-bins variance and the between-bins variance is computed.
  - If the light curve is folded using its true period, the AoV statistic is expected to reach a minimum value.
- In SL methods, the light curve is folded using a trial period and the sum of distances between consecutive points (string) in the folded curve is computed.
  - The true period is estimated by minimizing the string length on a range of trial periods.
  - The true period is expected to yield the most ordered folded curve and hence the minimum total distance between points

Slotted autocorrelation and beyond...

- In slotted autocorrelation, time lags are defined as intervals or slots instead of single values. The slotted autocorrelation function at a certain time lag is computed by averaging the cross product between samples whose time differences fall in the given slot.

- All these methods are essentially based on second order statistic analysis.
  - Information theoretic based criteria extract information from the probability density function.

- The slotting technique was extended to the information theoretic concept of correntropy. However, the slotted technique has the drawback that is highly dependant on the slot size.

Generalized Correlation Function

- The generalized correlation function (GCF) is also known as correntropy. The GCF measures similarities between feature vectors separated by a certain time delay in input space.

- The similarities are measured in terms of inner products in a high-dimensional kernel space.

- For a random process $\{X_t, t \in T\}$ with T being an index set, the correntropy function is defined as:

$$V(t_1, t_2) = \mathbb{E}_{x_{t_1} x_{t_2}}[\kappa(x_{t_1}, x_{t_2})],$$

Centered Correntropy

- It is also possible to define the centered correntropy:

$$U(t_1, t_2) = \mathbb{E}_{x_{t_1} x_{t_2}}[\kappa(x_{t_1}, x_{t_2})] - \mathbb{E}_{x_{t_1}} \mathbb{E}_{x_{t_2}}[\kappa(x_{t_1}, x_{t_2})]$$

- A kernel can be viewed as a measure of the similarity of the data.

- A frequently used kernel is the Gaussian kernel:

$$G_\sigma(x - z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$$

Correntropy Kernelized Periodogram

- A kernel can be any semi-definite positive function. And it is possible to develop periodic kernels.

- This brings to a metric combining the correntropy with a periodic kernel to measure similarity among samples separated by a given period.

- This algorithm is known as Correntropy Kernelized Periodogram, and does not require any resampling, slotting or folding scheme as it is computed directly from the available samples.

- This idea has strong connection with Gaussian Process analysis, machine learning and information theory,

Mutual Information

- Another modern approach to the problem tries to measure the degree of dependence between different random variables (RVs).
  - MI is able to capture nonlinear dependence too.

- Technically speaking, the MI is the divergence (i.e. statistical distance) between the joint PDF of the RVs and the product of their marginal PDFs.

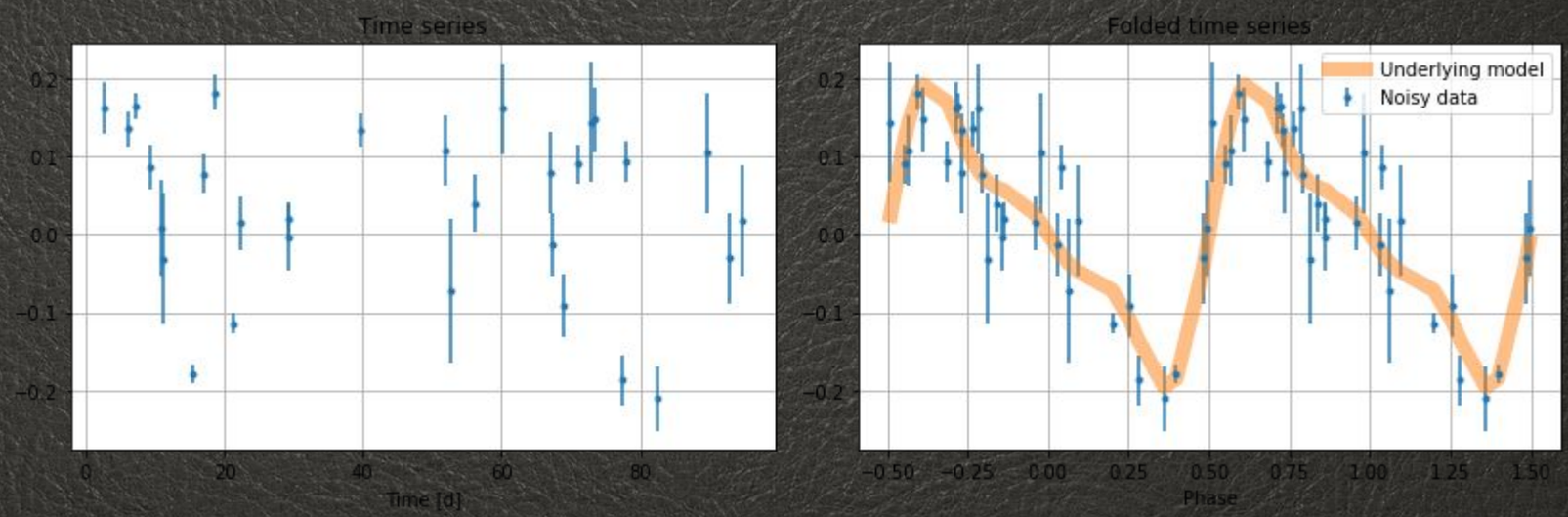
$$\begin{aligned} \mathrm{MI}_S(X, Y) &= D_{KL}(f_{X,Y} \| f_X f_Y) \\ &= \iint f_{X,Y} \log f_{X,Y} \, dx \, dy \\ &\quad - \int f_X \log f_X \, dx - \int f_Y \log f_Y \, dy, \end{aligned}$$

- where D_{KL} is the Kullback-Leibler divergence and f_X and f_Y are the marginal PDFs of X and Y, respectively.

Exercise

- Useful notebook:

    1. NonParametricPeriodograms



REFERENCES AND DEEPENING

THE ASTROPHYSICAL JOURNAL, 224:953–960, 1978 September 15

PERIOD DETERMINATION USING PHASE DISPERSION MINIMIZATION

R. F. STELLINGWERF
Department of Physics and Astronomy, Rutgers University
Received 1978 February 6; accepted 1978 March 22

ABSTRACT

We derive a period determination technique that is well suited to the case of nonsinusoidal time variation covered by only a few irregularly spaced observations. A detailed statistical analysis allows comparison with other techniques and indicates the optimum choice of parameters for a given problem. Application to the double-mode Cepheid BK Cen demonstrates the applicability of these methods to difficult cases. Using 49 photoelectric points, we obtain the two primary oscillatory components as well as the principal mode-interaction term; the derived periods are in agreement with previous estimates.

Subject headings: stars: Cepheids — stars: individual — stars: pulsation

Bob Stellingwerf

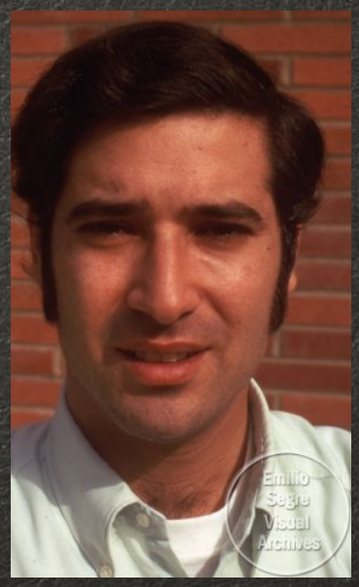
Mon. Not. R. astr. Soc. (1983) **203**, 917–924

A period-finding method for sparse randomly spaced observations or "How long is a piece of string?"

M. M. Dworetsky *Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT*

Received 1982 August 17; in original form 1982 March 26

Summary. A string-length method for establishing the period of a variable star from a relatively small number of randomly spaced observations over a long span of time is investigated. Criteria for establishing the validity of indicated periods are presented. The method is particularly suited for determination of periods in the limiting case of relatively few observations of reasonably high accuracy. A revised period and orbital elements for a spectroscopic binary observed by Abt & Levy (17 Lyr) are given. The period given by Abt & Levy for 18 Com is not confirmed; the data are insufficient to determine the correct period.



Mike Dworetsky